MAXIMUM ENTROPY RELAXATION FOR GRAPHICAL MODEL SELECTION GIVEN INCONSISTENT STATISTICS

Venkat Chandrasekaran, Jason K. Johnson, Alan S. Willsky

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

We develop a novel approach to approximate a specified collection of marginal distributions on subsets of variables by a globally consistent distribution on the entire collection of variables. In general, the specified marginal distributions may be inconsistent on overlapping subsets of variables. Our method is based on maximizing entropy over an exponential family of graphical models, subject to divergence constraints on small subsets of variables that enforce closeness to the specified marginals. The resulting optimization problem is convex, and can be solved efficiently using a primal-dual interior-point algorithm. Moreover, this framework leads naturally to a solution that is a sparse graphical model.

Index Terms— Graphical models, maximum entropy principle, model selection, inconsistent statistics

1. INTRODUCTION

Graphical models provide a powerful framework for statistical signal processing. They offer a convenient representation for joint probability distributions and convey the Markov structure in a large number of random variables compactly [1]. A graphical model is a collection of variables defined with respect to a graph; each vertex of the graph is associated with a random variable and the edge structure specifies the conditional independence (Markov) properties among the variables. In many problems, the Markov structure underlying a collection of variables $x_V = \{x_v | v \in V\}$ is not known and must be learned from empirical observations of the variables. In areas of the sciences such as geophysics, medicine, and oceanography, one often only has access to marginal empirical data samples for subsets of variables x_{V_k} ($V_k \subset V$) separately, rather than for the entire collection x_V jointly. The subsets V_k may not be mutually disjoint in general, thus leading to marginal empirical statistics that can be *inconsistent* on overlapping subsets of variables.

In this paper, we develop a novel approach to approximate a specified collection of marginal distributions on subsets of variables by a sparse, *globally consistent* graphical model on the entire collection of variables. We assume that each variable is observed as part of at least one of the subsets V_k so that $\cup_k V_k = V$. Our method is based on maximizing entropy subject to marginal divergence constraints on small subsets of variables. The marginal divergence constraints are constructed in a manner that takes into account the varying degrees of confidence for the different specified marginal statistics. When appropriately viewed in the context of exponential families, our formulation reduces to a convex optimization program that can be efficiently solved using a primal-dual interior-point algorithm. If the data samples are observed on the entire collection of variables jointly, we recover the framework in [2], which deals with the case where the specified statistics are globally consistent. We note that our framework for model selection is also applicable when the empirical observations contain data values missing at random [3].

In Section 2, we provide a brief background on graphical models and exponential families. We discuss our maximum-entropy relaxation framework in Section 3. In Section 4, we demonstrate the effectiveness of our approach in learning the structure of simple graphical models given inconsistent statistics. Our framework is applicable for both Gaussian and discrete graphical model selection; however, the simulation results focus exclusively on Gaussian models with experimental examples for discrete models deferred to a longer paper. We conclude with a brief discussion in Section 5.

2. BACKGROUND

2.1. Graphical models and Exponential families

A *graphical model* [1] is a collection of random variables indexed by the vertices of a graph $\mathcal{G} = (V, \mathcal{E})$; each vertex $v \in V$ corresponds to a random variable x_v, and where for any $A \subset V$, $x_A = \{x_v | v \in A\}$. The set $\mathcal{E}$ is some subset of $\binom{V}{2}$, the set of all pairs of edges[1]. A subset $S \subset V$ is said to *separate* $A, B \subset V$ if every path between a vertex in A and

[1]This notion can be generalized to include high-order edges involving more than two variables.

one in B passes through a vertex in S. A distribution $p(x_V)$ is *Markov* with respect to a graph $\mathcal{G} = (V, \mathcal{E})$ if for any subsets $A, B \subset V$ that are separated by some $S \subset V$, the subset of variables x_A is conditionally independent of x_B given x_S, i.e. $p(x_A, x_B|x_S) = p(x_A|x_S) \cdot p(x_B|x_S)$.

A distribution being Markov with respect to a graph implies that it can be decomposed into local functions in a very particular way [1]. We elaborate on this connection for exponential family distributions [4]. Let $\mathbb{X}$ be either a continuous or discrete sample space. We consider parametric families of probability distributions with support $\mathbb{X}^{|V|}$ defined by

$$p_\theta(x) = \exp\{\theta^T \phi(x) - \Phi(\theta)\}, \tag{1}$$

where $\phi : \mathbb{X}^{|V|} \to \mathbb{R}^d$ are the *sufficient statistics*, θ are the *exponential parameters*, and $\Phi(\theta) = \log \int \exp(\theta^T \phi(x)) dx$ is the *log-partition function*.[2] The family is defined by the set $\Theta \triangleq \left\{\theta \in \mathbb{R}^d : \Phi(\theta) < \infty\right\} \subset \mathbb{R}^d$ of all normalizable θ. A class of graphical models is obtained by defining the collection of statistics ϕ to be local functions over small subsets of variables. Let $\phi = \{\phi_v(x_v), v \in V\} \cup \{\phi_E(x_E), E \in \binom{V}{2}\}$ define a collection of node and pairwise statistics, where each $\phi_E(x_E)$ (or $\phi_v(x_v)$) is only a function of the variables x_E (or variable x_v). Specializing the Hammersley-Clifford theorem [1] to such exponential family distributions, we have that if p_θ is Markov with respect to $\mathcal{G} = (V, \mathcal{E})$, then θ is sparse according to $\mathcal{G}$, i.e. $\theta_E = 0$ for $E \notin \mathcal{E}$. The set of distributions $\Theta(\mathcal{G}) \subset \Theta$ that are Markov with respect to a graph $\mathcal{G} = (V, \mathcal{E})$ can also be viewed as an exponential family with the reduced set of statistics $\{\phi_v(x_v), v \in V\} \cup \{\phi_E(x_E), E \in \mathcal{E}\}$, and the corresponding reduced set of exponential parameters $\{\theta_v, v \in V\} \cup \{\theta_E, E \in \mathcal{E}\}$.

By taking expectations of the statistics with respect to $p_\theta(x)$, we obtain the *moment parameters*

$$\eta = \mathbb{E}_{p_\theta}\left\{\phi(x)\right\}. \tag{2}$$

Let $\mathcal{M}$ denote the set of *realizable* moment parameters that can be obtained under expectations with respect to some $\theta \in \Theta$. For a linearly independent set of statistics ϕ, (2) defines a bijective map $\Lambda : \Theta \to \mathcal{M}$ from exponential to moment parameters, thus allowing an alternate moment parameterization of an exponential family distribution. Distributions parameterized by the moment parameters are denoted by $p_\eta(x) \triangleq p_{\Lambda^{-1}(\eta)}(x)$. The set of distributions $\mathcal{M}(\mathcal{G}) \subset \mathcal{M}$ that are Markov with respect a graph $\mathcal{G} = (V, \mathcal{E})$ are parameterized by the subset of moment parameters $\{\eta_v, v \in V\} \cup \{\eta_E, E \in \mathcal{E}\}$.

2.2. Entropy and Divergence

The entropy of an exponential family distribution parameterized by the moment parameters is the conjugate [5] of the log-partition function:

$$H(p_\eta(x)) \triangleq H(\eta) = \min_{\theta \in \Theta} \Phi(\theta) - \eta^T \theta.$$

[2]The integral must be replaced by a sum for discrete models.

The entropy function is concave as a function of the moment parameters. The gradient of entropy is given by the negative of the corresponding exponential parameters:

$$\nabla_\eta H(\eta) = -\Lambda^{-1}(\eta).$$

Further, the Hessian $\nabla^2_\eta H(\eta)$ with respect to η is given by the negative Fisher information matrix $-G(\eta)$ with respect to the moment parameters η, where $G(\eta)$ is given by:

$$G(\eta) \triangleq \mathbb{E}_{p_\eta(x)}\left[(\nabla_\eta \log p_\eta(x))(\nabla_\eta \log p_\eta(x))^T\right].$$

Finally, we note that the Kullback-Leibler divergence [6] between two distributions parameterized by moment parameters $D(p_\eta(x) \| p_\nu(x)) \triangleq D(\eta \| \nu)$ is the *Bregman* distance induced by the entropy function:

$$D(\eta \| \nu) = \left[H(\nu) + (\eta - \nu)^T \nabla_\nu H(\nu)\right] - H(\eta).$$

Thus, $D(\eta \| \nu)$ is convex with respect to the moment parameters η, keeping the moments ν fixed. The gradient and Hessian are given as follows:

$$\begin{aligned} \nabla_\eta D(\eta \| \nu) &= \Lambda^{-1}(\eta) - \Lambda^{-1}(\nu) \\ \nabla^2_\eta D(\eta \| \nu) &= G(\eta). \end{aligned}$$

2.3. Gaussian models as an exponential family

Let $\{x_v | v \in V\}$ be a zero-mean[3] Gaussian graphical model with a symmetric positive-definite covariance matrix P [1]. A natural parameterization for such a model that provides a connection to exponential families is in terms of the *information matrix* $J = P^{-1}$, so that $p(x_V) \propto \exp\left\{-\frac{1}{2} x_V^T J x_V\right\}$. Thus, if $p(x_V)$ is Markov with respect to $\mathcal{G} = (V, \mathcal{E})$, then $J_{v,u} = J_{u,v} = 0$ if and only if the edge $\{v, u\} \notin \mathcal{E}$ for every pair of vertices $v, u \in V$. Defining statistics $\phi_v(x_v) = x_v^2, \forall v \in V$, and $\phi_{v,u}(x_v, x_u) = x_v x_u, \forall \{v, u\} \in \binom{V}{2}$, we obtain θ and η parameters that are respectively given by elements of the J and P representations:

$$\theta = \left(-\tfrac{1}{2} J_{vv}, \forall v\right) \cup \left(-J_{vu}, \forall \{v, u\}\right) \tag{3}$$

$$\eta = (P_{vv}, \forall v) \cup (P_{vu}, \forall \{v, u\}). \tag{4}$$

A key point here is that the marginal density for a subset of variables is determined by the corresponding subset of the moment parameters (a principle submatrix of P).

The entropy and divergence in Gaussian models parameterized by covariance matrices P, Q (i.e., moment parameters) is given by [6]:

$$\begin{aligned} H(P) &= \tfrac{1}{2}(\log\det P + |V| \log 2\pi e) \\ D(P \| Q) &= \tfrac{1}{2}(\mathrm{tr}(PQ^{-1}) - \log\det PQ^{-1} - |V|). \end{aligned}$$

The Fisher information matrix $G(\eta)$ is given by $G_{vu,ij} = J_{vi} J_{uj} + J_{vj} J_{ui}$, $G_{vu,i} = J_{vi} J_{ui}$, and $G_{v,u} = \frac{1}{2} J^2_{vu}$, where $J = P(\eta)^{-1}$ [7].

[3]The mean vector does not play a significant role in model selection.

2.4. Maximum entropy principle

The maximum-entropy principle [8] states that subject to linear constraints on a set of statistics, the entropy-maximizing distribution among *all* distributions lies in the exponential family based on those statistics used to define the constraints. Consider the following restricted maximum-entropy problem within the framework of exponential family distributions [4]. Let η be the moment parameters of an exponential family, and let η_V and $\eta_{\mathcal{E}}$ represent the subset of moment parameters corresponding to the set of all vertices V and a set of edges $\mathcal{E}$ respectively. We constrain these moments to be equal to η_V^* and $\eta_{\mathcal{E}}^*$ (for example, these could be empirical moments):

$$\text{(ME)} \quad \begin{array}{ll} \arg\max_{\eta \in \mathcal{M}} & H(\eta) \\ \text{s.t.} & \eta_{\mathcal{E}} = \eta_{\mathcal{E}}^*, \eta_V = \eta_V^*. \end{array}$$

Based on the maximum-entropy principle, we can conclude that the optimal distribution (if it exists) of this ME problem over the *entire* exponential family $\{p_{\Lambda^{-1}(\eta)} : \eta \in \mathcal{M}\} = \{p_\theta : \theta \in \Theta\}$ is Markov with respect to the graph $(V, \mathcal{E})$. That is, letting $\tilde{\eta}$ denote the solution of ME, we have that $\tilde{\eta} \in \mathcal{M}(\mathcal{G})$. This suggests that entropy, when used as a maximizing objective function, favors *sparse* graphical models.

3. PROBLEM FORMULATION AND MAXIMUM ENTROPY RELAXATION

Let $\{V_k\}$ be a finite collection of subsets of vertices such that $\cup_k V_k = V$. In general, the subsets $\{V_k\}$ are not mutually disjoint. For each V_k, we are given N_k independent, identically distributed observations only of the variables x_{V_k}. Hence, the overall collection of observations can be specified as $\{V_k, \{x_{V_k}^i\}_{i=1}^{N_k}\}$. Consider subsets V_k and $V_{k'}$ that are not disjoint. The marginal empirical statistics of the variables $x_{V_k \cap V_{k'}}$ obtained from $\{x_{V_k}^i\}_{i=1}^{N_k}$ and $\{x_{V_{k'}}^i\}_{i=1}^{N_{k'}}$ are *inconsistent* with each other. The goal is to find a globally consistent, sparse graphical model approximation that is still close to the empirical marginal statistics obtained from the samples $\{V_k, \{x_{V_k}^i\}_{i=1}^{N_k}\}$. We note that if the observations are over the entire collection of variables jointly so that we have samples $\{x_V^i\}_{i=1}^N$, the empirical statistics are globally consistent.

3.1. Motivation

Before presenting our framework, we briefly discuss previously studied approaches to this problem based on maximizing the likelihood function [3, 9]. When the observations are over the entire collection of variables jointly $\{x_V^i\}_{i=1}^N$, maximizing the log-likelihood leads to a convex optimization problem and the resulting estimate is the same as that produced by the ME approach of Section 2.4. However, given observations $\{V_k, \{x_{V_k}^i\}_{i=1}^{N_k}\}$ the log-likelihood function

$$\ell(\theta) = \sum_{V_k} \sum_{i=1}^{N_k} \log p_\theta(x_{V_k}^i),$$

with $p_\theta(x_{V_k})$ being the marginal distribution over x_{V_k}, is non-convex due to the marginalization [9]. Thus, the EM algorithm [9] does not always converge to the maximum likelihood estimator, but rather to other local maxima.

Our framework provides a *convex* model selection procedure based on a relaxation of the ME approach. By virtue of the fact that our optimization problem involves maximizing entropy, it identifies a sparse graphical model approximation that is globally consistent. Further, we note that the estimates produced by our approach are parametrically consistent (i.e., converge to the correct underlying model parameters) in the large-sample limit under very mild conditions.

3.2. Maximum entropy relaxation

Motivated by the maximum-entropy principle, we propose the following *relaxed* maximum-entropy formulation:

$$\text{(MER)} \quad \begin{array}{ll} \arg\max_{\eta \in \mathcal{M}} & H(\eta) \\ \text{s.t.} & D_E(\eta_{\bar{E}} \| \eta_{\bar{E}}^*) \leq \delta_E, \ \forall E \in \mathcal{E} \\ & D_v(\eta_v \| \eta_v^*) \leq \delta_v, \ \forall v \in V. \end{array}$$

Here, D_E and D_v are the marginal divergences on $E \in \mathcal{E}$ and $v \in V$ respectively, the edge set $\mathcal{E}$ serves to specify the constraint set, and $\delta = \{\delta_E, E \in \mathcal{E}\} \cup \{\delta_v, v \in V\}$ are a specified set of tolerances on marginal divergences. The moments $\eta_{\bar{E}}$ refer to the collection of moments that have support inside edge E, and specify the moments of the marginal distribution of variables x_E; for example, $\eta_{\{v,u\}} = \{\eta_v, \eta_u, \eta_{vu}\}$. The moments $\eta_{\bar{E}}^*$ and η_v^* denote marginal empirical statistics on edge E and vertex v respectively. These statistics are based on the empirical data samples $\{V_k, \{x_{V_k}^i\}_{i=1}^{N_k}\}$; we define them in Section 3.3.

If the observations are over the entire collection of variables jointly $\{x_V^i\}_{i=1}^N$ (leading to empirical statistics that are globally consistent), the MER problem is feasible for all $\delta \geq 0$ and reduces to the formulation in [2]. However, given observations $\{V_k, \{x_{V_k}^i\}_{i=1}^{N_k}\}$ that lead to inconsistent empirical statistics, the tolerances must be large enough so that the constraint set is non-empty. In this case, MER is not feasible for all $\delta \geq 0$. We discuss this issue in detail in Section 3.4 where we provide a method to choose δ in order to ensure feasibility.

When the statistics ϕ in the exponential family are linearly independent, the entropy $H(\eta)$ is a strictly concave function of η. The set of realizable moments $\mathcal{M}$ is a convex subset of $\mathcal{R}^d$, and each marginal divergence $D_E(\eta_{\bar{E}} \| \eta_{\bar{E}}^*)$ (or $D_v(\eta_v \| \eta_v^*)$) is a convex function of $\eta_{\bar{E}}$ (or η_v) for any fixed value of $\eta_{\bar{E}}^*$ (or η_v^*). Hence, this is a convex optimization problem [10]. Thus, if the maximum entropy is obtained by some $\tilde{\eta} \in \mathcal{M}$, it is the unique solution of the MER problem.

Note that we have not imposed any restrictions on the set of edges $\mathcal{E}$, and in general, this set could even be the complete set $\binom{V}{2}$. However, we expect that the entropy objective would favor a sparse Markov model that lies within the constraint set, with the degree of sparsity obtained in the solution

being controlled by the tolerances δ. The following theorem precisely illuminates the model-thinning property of the MER problem. Let $\mathcal{E}_{active} \subseteq \mathcal{E}$ denote the set of edge constraints that are satisfied with equality (also called active edge constraints) by the solution of the MER problem.

Theorem 1 *The solution of the MER problem (if it exists) is Markov with respect to the graph $(V, \mathcal{E}_{active})$ defined by the active edge constraints. (Thus, the solution is also Markov with respect to the graph $(V, \mathcal{E})$ defined by all the constraints.)*

Proof: Based on the Karush-Kuhn-Tucker conditions and complementary slackness [10]. See [2] for more details. □

3.3. Defining constraints

We are provided with observations $\{V_k, \{x^i_{V_k}\}_{i=1}^{N_k}\}$. Define $\mathcal{E}_k \triangleq \binom{V_k}{2}$ to be the set of all pairwise edges corresponding to V_k. Let $\mathcal{E} = \cup_k \mathcal{E}_k$ in the MER problem, so that the edge constraints are only defined for those edges for which both variables are jointly observed. For each node $v \in V$, consider only those subsets from $\{V_k\}$ that contain v. Let $N_v = \sum_{k:v\in V_k} N_k$. We can then compute the empirical statistic η^*_v as follows:

$$\eta^*_v = \frac{\sum_{k:v\in V_k} \sum_{i=1}^{N_k} \phi_v((x^i_{V_k})_v)}{N_v}. \tag{5}$$

Let $\phi_{\bar{E}}(x_E)$ be the collection of statistics that are functions of variables inside edge E. This definition is analogous to that of the moment parameters $\eta_{\bar{E}}$ so that $\eta_{\bar{E}} = \mathbb{E}_{p_\theta}\{\phi_{\bar{E}}(x_E)\}$. As with the vertices, let $N_E = \sum_{k:E\in\mathcal{E}_k} N_k$. For each edge $E \in \mathcal{E}$, we define the empirical statistics $\eta^*_{\bar{E}}$ as follows:

$$\eta^*_{\bar{E}} = \frac{\sum_{k:E\in\mathcal{E}_k} \sum_{i=1}^{N_k} \phi_{\bar{E}}((x^i_{V_k})_E)}{N_E}. \tag{6}$$

These empirical statistics take into account every occurrence of the variables x_E together so that $\eta^*_{\bar{E}}$ can be defined in a coherent manner. Based on these definitions, the inconsistency arises in the node empirical statistics. More precisely, if a node v belongs to edge E, then in general $\eta^*_v \neq (\eta^*_{\bar{E}})_v$.

Consider a simple example with variables $\{x_1, x_2, x_3\}$. Suppose that we have N_{12} observations of variables $\{x_1, x_2\}$, N_{23} observations of variables $\{x_2, x_3\}$, and N_{13} observations of variables $\{x_1, x_3\}$. In defining η^*_1, we use the samples of x_1 from both the N_{12} observations of $\{x_1, x_2\}$ and the N_{13} observations of $\{x_1, x_3\}$. However, in defining $\eta^*_{\{1,2\}}$ we only use the N_{12} observations of $\{x_1, x_2\}$. Thus, $\eta^*_1 \neq (\eta^*_{\{1,2\}})_1$.

We note here that when the data samples are observed over the entire collection of variables jointly $\{x^i_V\}_{i=1}^N$, we do not have inconsistencies in the empirical statistics as defined in (5−6) and our framework reduces to the description in [2].

Specifying tolerances: For large sample sizes, the expectations of $D_v(\eta_v\|\eta^*_v)$ and $D_E(\eta_{\bar{E}}\|\eta^*_{\bar{E}})$ decay as $\frac{1}{N_v}$ and $\frac{3}{N_E}$ respectively [11]. The assumption here is that η_v and $\eta_{\bar{E}}$ are the (fixed) true moments, and η^*_v and $\eta^*_{\bar{E}}$ are empirical averages obtained from random samples drawn according to the true moments. The scaling factors differ because each vertex contains one parameter, while each edge contains three. Using this as motivation, we choose $\delta_v = \gamma(\frac{1}{N_v})^p$ and $\delta_E = \gamma(\frac{3}{N_E})^p$ with $p < 1$ to obtain a slower decay than $\frac{1}{N_v}$ and $\frac{3}{N_E}$. The parameter γ trades off between accuracy and complexity of the MER solution with larger γ leading to sparser graphical models. In practice, one could also employ cross-validation methods to guide the choices of δ.

3.4. Feasibility

We solve the MER problem by using a primal-dual interior-point algorithm as described in [2, 10]. An important first step in this method is to ensure that the MER problem is feasible (i.e., choose δ large enough), and to identify a strictly feasible point η_{init} (one that satisfies all the inequality constraints strictly) for use as an initial guess in the algorithm. Note that η_{init} must correspond to a globally consistent set of moments, i.e. $\eta_{init} \in \mathcal{M}$. When the observations are over the entire collection of variables jointly $\{x^i_V\}_{i=1}^N$, we can simply set η_{init} to be equal to the globally consistent empirical statistics defined in (5−6). In this case, the MER problem is feasible for all $\delta \geq 0$. However, given observations $\{V_k, \{x^i_{V_k}\}_{i=1}^{N_k}\}$ that lead to inconsistent empirical statistics, the MER problem is not feasible for all $\delta \geq 0$. Therefore, we solve the following optimization problem as a pre-processing step:

$$\text{(INI)} \quad \begin{array}{ll} \arg\min_{\eta\in\mathcal{M}(\mathcal{G}), s\geq 0} & s \\ \text{s.t.} & D_E(\eta_{\bar{E}}\|\eta^*_{\bar{E}}) \leq s(\frac{3}{N_E})^p, \forall E \in \mathcal{E} \\ & D_v(\eta_v\|\eta^*_v) \leq s(\frac{1}{N_v})^p, \forall v \in V. \end{array}$$

This problem is a modified version of a so-called basic phase-I method [10]. Note that the optimization is over $s \geq 0$ *and* $\eta \in \mathcal{M}(\mathcal{G})$. Letting the optimal value of INI be $\tilde{s}$, we are guaranteed that the MER problem is feasible for any $\gamma > \tilde{s}$. Letting the optimal moments produced by solving INI be η_{init}, we can use η_{init} as a feasible initial guess in solving MER for all $\gamma > \tilde{s}$. The INI optimization problem is convex, and can be solved efficiently using a barrier method [10].

4. EXPERIMENTAL RESULTS

We solve the MER problem using a primal-dual interior-point algorithm [2, 10]. We employ a boot-strapping approach that solves a sequence of tractable sub-problems and exploits the sparsity of the Fisher information matrix in order to compute successive primal-dual search directions efficiently (see [2] for details). Indeed, this boot-strapping approach could also be useful for solving the INI problem efficiently. In this section, we present two simple Gaussian model selection experiments with different data observation schemes, each leading to inconsistent empirical statistics.

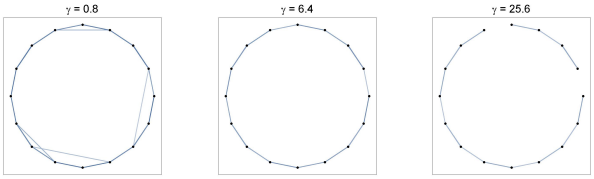

Fig. 1. 16-node cycle experiment: Graphs of the MER solution for various values of γ.

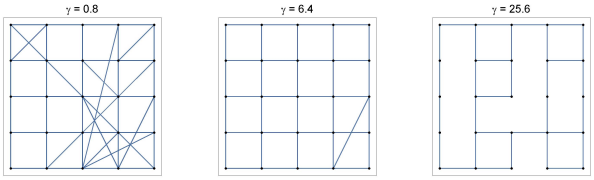

Fig. 2. 5x5 grid experiment: Graphs of the MER solution for various values of γ.

First, we consider samples drawn from a 16-node cyclic Gaussian model, where the nodes are arranged in a circle and edges connect nodes that are one step away on the circle. The node weights are $J_{vv} = -2\theta_v = 1.0$ for every node v, and the edge weights are $J_{uv} = -\theta_{uv} = -0.1875$ for each edge $\{u, v\}$. We generate 100 samples each for variable subsets $\{1, \ldots, 8\}$, $\{5, \ldots, 12\}$, $\{9, \ldots, 16\}$, and $\{1, \ldots, 4\} \cup \{13, \ldots, 16\}$. Figure 1 shows the MER solution graphs for $p = 0.5$ and various values of γ. Notice that as the value of γ increases, the effect of the relaxation is stronger and fewer edges are included in the MER solution. For $\gamma = 6.4$ the correct underlying graph structure is recovered.

Next, we consider 500 samples drawn from a 5x5 nearest-neighbor grid-structured model where data values are missing with probability 0.2. That is, in every sample each variable has a 20% chance of being unobserved, independent of the value of the variable [3]. The node weights in the underlying model are $J_{vv} = -2\theta_v = 1.0$ for every node v, and the edge weights are $J_{uv} = -\theta_{uv} = -0.24$ for each edge $\{u, v\}$. Figure 2 shows the MER solution graphs for $p = 0.5$ and various values of γ. The underlying graph structure is recovered for $\gamma = 6.4$ with very few spurious or missing edges.

5. DISCUSSION

We describe a framework for model selection given observations of subsets of a collection of variables separately, rather than of the entire collection jointly. In general, the subsets are not disjoint, thus leading to inconsistent marginal empirical statistics on overlapping subsets of variables. Our formulation maximizes entropy subject to marginal divergence constraints on small subsets of variables. Viewed within the class of exponential families, the optimization problem is convex, and leads naturally to a globally consistent, sparse graphical model approximation to the empirical statistics. Our framework is also applicable to the commonly-encountered scenario in which data values within each sample are missing at random [3].

We envision several directions for future research. Our framework may be applicable in problems where one only has access to empirical moments and "confidences" for these moments [12]. Finally, a detailed generalization error analysis of our approach would be useful for choosing the parameters p and γ in the tolerances.

6. REFERENCES

[1] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, U.K., 1996.

[2] J. K. Johnson, V. Chandrasekaran, and A. S. Willsky, "Learning Markov Structure by Maximum Entropy Relaxation," in *Eleventh International Conference on Artificial Intelligence and Statistics*, Puerto Rico, 2007.

[3] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley Interscience, New Jersey, 2002.

[4] S. Amari, "Information geometry on a hierarchy of probability distributions," *IEEE Transactions on Information Theory*, vol. 47, no. 5, 2001.

[5] R. T. Rockafellar, *Convex Analysis*, Princeton University press, 1996.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 1991.

[7] J. K. Johnson, *Relaxation Methods for Learning and Inference in Graphical Models*, Ph.D. thesis, Massachusetts Institute of Technology, 2007, In preparation.

[8] E. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 16, no. 4, 1957.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, 1977.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[11] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, 1974.

[12] J. B. Predd, S. R. Kulkarni, D. N. Osherson, and H. V. Poor, "Scalable Algorithms for Aggregating Disparate Forecasts of Probability," in *Ninth International Conference on Information Fusion*, Florence, Italy, 2006.